ICIC Express Letters
Volume **11**, Number **3**, March **2017**
 ISSN 1881-803X
pp. **625–632**

HOT TOPIC DETECTION AND ANALYSIS ON TEMPORAL MICROBLOG TOPIC MODEL

Mei Yu[1,2], Hongyun Shang[1,3], Jian Yu[1,2], Tianyi Xu[1,2,*]
Jie Gao[1,2] and Yue Gao[2]

[1]Tianjin Key Laboratory of Cognitive Computing and Application
[2]School of Computer Science and Technology
[3]School of Software
Tianjin University
No. 92, Weijin Road, Nankai District, Tianjin 300072, P. R. China
yumei@tju.edu.cn; Shanghongyun123@outlook.com; *Corresponding author: tianyi.xu@tju.edu.cn

Received July 2016; accepted October 2016

Abstract. *Detecting hot topics from social network is a significant problem on Internet public opinion management. Most existing topic detection algorithms are based on n-grams clustering techniques or latent topic detection, which lack the ability to detect topics in an "on-line" manner, since they ignore the temporal aspect of the topics they aim to track. In this paper, a temporal microblog topic model for detecting hot topics in microblogs is presented based on the fact that temporal changes exist in all topics. Therefore, the whole period is discretized to several temporal series. The LDA model is optimized by combining with the temporal character to detect topics in all temporal series. Then, Jensen-Shannon divergence is chosen to compute distance of the K-means clustering algorithm, which is applied to finding hot topics. Finally, a hot topic popularity trend diagram about microblog is established and trend analyses are provided. The proposed hot topic detection technique is compared with LDA model and random selection model. The TMT model outperformed these non-temporal algorithms with an improved accuracy of topic detection and an excellent performance on hot topic clustering. This would provide effective basis for hot public opinion trend analysis.*
Keywords: Internet public opinion, Hot topic detection, Temporal information, Microblog

1. **Introduction.** Microblog has become an important information source of hot topics and emergencies. As demonstrated in [1-3], microblog hot topic detection could be used for analyzing the evolution of social phenomena over time, obtaining information on people's opinion about an event, or improving situational awareness and the impact of public policies.

Most existing topic detection algorithms are based on latent Dirichlet allocation (LDA) model. LDA model is an unsupervised machine learning technology introduced by Blei et al. in [4]. It is generally used to recognize latent topic information in large scale of document collection or corpus. These topic detection researches are all based on the assumption that the subject of off-line content generated by microblogs changes little in the whole period. However, this assumption is not exact, because each topic would have a life cycle containing the stages such as appearance, development and disappearance. A shorter duration latent topic would be covered by a longer duration topic without considering the temporal information. Taking the temporal character into account, the accuracy of topic detection for microblogs would be increased.

In this paper, an efficient algorithm named temporal microblog topic (TMT) model is presented for detecting hot topics in microblog. The TMT model is based on the fact that temporal changes exist in all topics. Temporal character is added to the LDA modeling

process, and appropriate distance computation is applied on K-means clustering, thereby an intuitive framework for hot topic detection in microblog is created. Our method could work in an "on-line" manner, also offer certain advantages over previous methods in terms of the accuracy of topic detection and the clustering result.

The subsequent paper structure is as follows. Section 2 summarizes the research status on topic detection. Section 3 describes the framework and the concrete theory of the TMT model. Section 4 introduces the process of experiment and a manifold performance evaluation. Section 5 is dedicated to conducting remarks and future work.

2. **Related Work.** Most state-of-the-art approaches to topic detection are based on n-grams clustering techniques or latent topic detection. LDA is a three-layer Bayes model, consisting of structures of word, topic and documentation. Many algorithms based on the LDA are proposed to detect topic information. A BiTerm LDA algorithm in [5] extracts consecutive word pairs for detecting relevant topics. [6] proposes a mixture Gaussian model for bursty word extraction so as to detect real-time event in twitter, but it ignores that not all topics are characterized as "bursty words". After all, different topics have different life cycles. [7] adopts a single-pass clustering technique by using latent Dirichlet allocation (LDA) model to extract the hidden microblog topics information. [8] achieves encouraging results for event detection by using aggressive filtering and hierarchical clustering. However, these methods ignore the "temporal" character of topic they aim to track, so they can only work in an "off-line" manner, and the accuracy of the detection results will be weakened.

Recently, temporal topic mining in microblogs is applied in only a small number of articles. [9] develops an incremental clustering framework to detect new topics, and employs a range of content as well as temporal features to help detect hot emerging topics promptly. [10] presents a method for detecting events in microblogs, based on the similarity of the related temporal series. Among these ideas, temporal and frequency components are considered adequately in topic detection, but lexical itself is ignored. Not all the words with a similar temporal behavior are semantically related, so temporal similarity could not be a unique measure of topic relatedness.

In this paper, the TMT model comprehensively considers temporal character and semantic information of lexicology, which could detect topics in an "on-line" manner with an improved accuracy. What is more, the hot topics resulted from clustering algorithm are more precise, since the sparseness of clustering data is weaken indirectly by TMT model.

3. **Temporal Microblog Topic Detection and Analysis Method.** The essential idea of TMT model is that topic has temporal feature. The TMT model is shown in Figure 1.

After basic preprocessing work on the microblogs collection, the TMT model should be executed. The main steps of TMT model are simply illustrated as follows: first, the whole period considering the microblogs data is discretized to several temporal series, and a new model combining LDA and temporal character is built to detect topics in all temporal series; then, a clustering algorithm with appropriate distance computation is applied on the detected topics to identifying hot topics; finally, the evaluation of hot topic popularity over time is obtained.

3.1. **Topic detection.** Each microblog is treated as a bag-of-words document, and then the contents of all microblogs can be expressed as a sequence of keyword collections in n temporal windows. M_t denotes a temporal document-keyword matrix in the tth temporal window, where $M_t \in R^{N_u \times N_w}$, and N_u and N_w are the numbers of documents and keywords, respectively. Each row of M_t represents the keyword counts for a particular

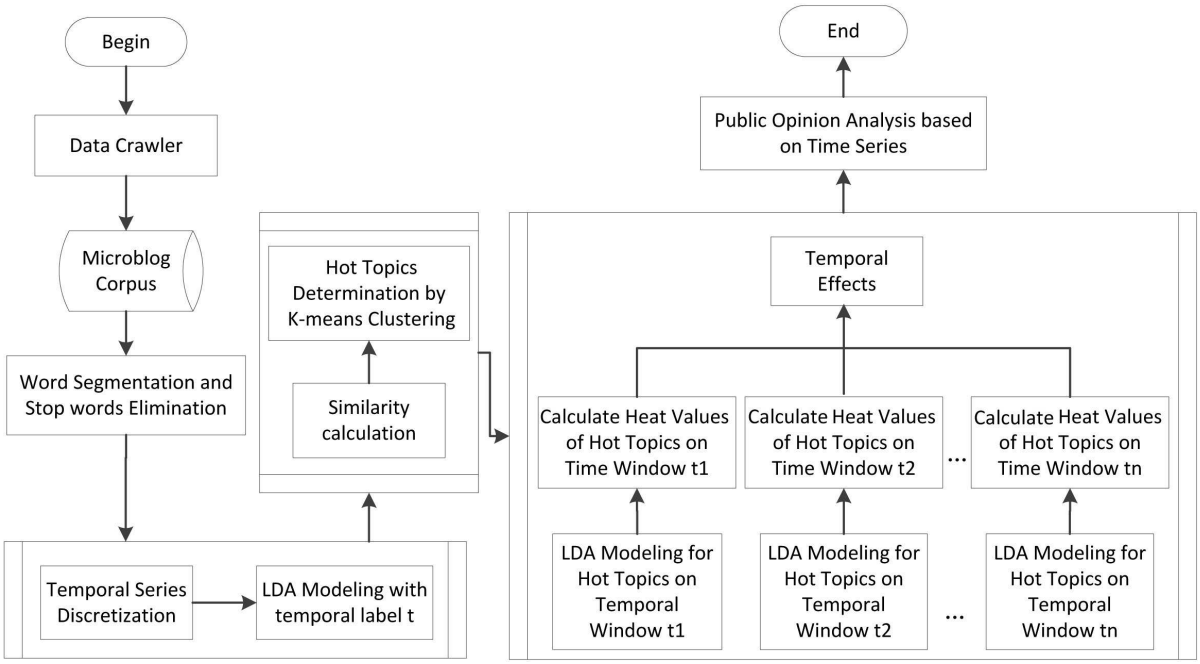

FIGURE 1. Process of the TMT model

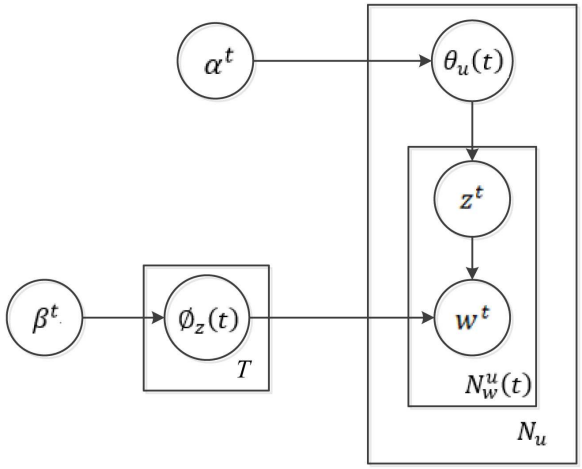

FIGURE 2. Generation of t-LDA model

document in the tth temporal window, whereas each column of M_t represents the counts of documents that contain a certain word in the tth temporal window.

To find topics of each document in M_t, the LDA model combined with temporal label t (t-LDA) is applied. Each document is associated with multiple topics, and each topic is represented by a probabilistic distribution over keywords. Formally, in the tth temporal window, the collection of N_u documents follows a multinomial distribution over T topics, which is denoted as $\theta_u(t)$; each topic follows a multinomial distribution over keywords, denoted as $\phi_z(t)$. $\theta_u(t)$ and $\phi_z(t)$ are respectively with Dirichlet prior parameters α^t and β^t. To get a keyword of document u, a topic z^t is sampled from the multinomial distribution $\theta_u(t)$ which is associated with document u in the tth temporal window, and then a keyword w^t corresponding to topic z^t is sampled from the multinomial distribution $\phi_z(t)$. This generation process is repeated $N_w^u(t)$ times to form keyword collection for document u.

The generation process of t-LDA model is shown in Figure 2.

Based on the model above, for the tth temporal window, the algorithm of extracting topic z^t from M_t is shown as Table 1.

TABLE 1. Algorithm in the tth temporal window

Algorithm 1 The t-LDA model in the tth temporal window

Input: document-keyword matrix M_t, Dirichlet prior parameters α^t and β^t
Output: topic sequence $z^t = \{z_1^t, z_2^t, \ldots, z_T^t\}$
Begin
  for topic $z^t = 1$, $z^t \leq T$, $z^t \leftarrow z^t + 1$
    Get a multinomial distribution of characteristic words, namely $\phi_z(t)$, β^t as Dirichlet prior.
  for document $u = 1$, $u \leq N_u$, $u \leftarrow u + 1$
    Get a multinomial distribution of topics, namely $\theta_u(t)$, α^t as Dirichlet prior.
    for word token $i = 1$, $i \leq N_w^u t$, $i \leftarrow i + 1$
      in the condition of u_i, sample a topic z_i^t from multinomial distribution $\theta_u(t)$
    in the condition of z_i^t, sample a word w_i^t from multinomial distribution $\phi_z(t)$
End

The topic-word distribution $\phi_z(t)$ and document-topic distribution $\theta_u(t)$ are estimated using the Gibbs sampling method mentioned in [11] as Formulas (1) and (2).

$$\phi_z(t) = \frac{C_{mj}^{WT} + \beta^t}{\sum_{m'=1}^{N_W^U(t)} C_{m'j}^{WT} + N_W^U(t)\beta^t} \tag{1}$$

$$\theta_u(t) = \frac{C_{hj}^{UT} + \alpha^t}{\sum_{j'=1}^{T} C_{hj'}^{UT} + T\alpha^t} \tag{2}$$

Among them, C^{WT} denotes the word-topic matrix with the scale of $N_w \times T$, and C_{mj}^{WT} denotes the times that word m is allocated to topic j. Similarly, C^{UT} denotes the document-topic matrix with the scale of $N_u \times T$, and C_{hj}^{UT} denotes the times that document h is allocated to topic j. By applying the t-LDA model to the entire keyword collections, topics for all documents in the temporal series are obtained.

3.2. **Hot topics determination.** Considering the content nondeterminacy of the detected topics, the K-means clustering is chosen to determine the potential hot topics. Hot topics are drawn from the clustering result. The distance between topic z^{t_i} and z^{t_j} is calculated by Jensen-Shannon divergence as Formula (3).

$$dist\left(z^{t_i}, z^{t_j}\right) = JSD(p,q) = \frac{1}{2}(D(p\|m) + D(q\|m)) \tag{3}$$

Suppose $i \leq j$, the topic z^{t_i} and z^{t_j} are extracted in temporal window t_i and t_j separately by the t-LDA model. The thesaurus is denoted by M; q is the probability distribution of z^{t_i} on the thesaurus M; p represents the probability distribution of z^{t_j} on the thesaurus M; $m = \frac{1}{2}(p+q)$. The Kullback-Leibler divergence (KL divergence) between p and q could represent the distribution difference between topic z^{t_i} and z^{t_j} on thesaurus, which is computed as Formula (4).

$$D(p\|q) = \sum_{i}^{|M|} p_i \log \frac{p_i}{q_i} \tag{4}$$

The smaller the difference is, the semantically closer two topics are. Note that the semantic relevance of two topics should be symmetric. So ultimately Jensen-Shannon divergence is chosen to overcome the asymmetry of KL divergence distance, which has been shown in Formula (3).

3.3. **Heat values calculation.** Here a temporal heat value model is built to achieve the evaluation of hot topics popularity over time. Based on the temporal series divided before, LDA model is built for each hot topic in every temporal window separately to extract characteristic words; thus, the document distribution involving each hot topic in every temporal window is clear. The topic heat value is reflected by the correlation between topic and document, which is calculated as Formula (5).

$$\delta_j^t = \frac{1}{D^t} \sum_{d \in D^t} \theta_{d,j} \tag{5}$$

The heat value of topic j in temporal window t is denoted as δ_j^t which represents popularity of this topic. $\theta_{d,j}$ expresses whether a document d in temporal window t belongs to topic j or not, and D^t is the number of documents in temporal window t. The heat value of a topic appropriately represents its popularity during a certain period. A hot topic popularity trend diagram could be established by organizing these heat values on the time line.

4. Experiment and Evaluation.

4.1. **Experimentation.** The data collection contains 25495 microblogs which spans from September 2011 to October 2011. Each microblog consists of publication time, published author, title, text and other information. The data set is segmented and then the stop words are removed only with nouns and verbs retained. Considering the time span of the data set, in order to guarantee that topics in every temporal window cover effective subject information sufficiently, 15 days are set as a unit, and thus the whole period is divided into 4 temporal windows.

The t-LDA model is used to extract characteristic words for each temporal window. Among the LDA model parameters, K is the total number of topics 150, which is set as 150 in this experiment. The estimation method of $\phi(t)$ and $\theta(t)$ is Gibbs sampling method which is used for finding the solution of the text vector matrix after modeling. The Gibbs sampling parameters are set as $\alpha^t = 50/T$ and $\beta^t = 0.01$, iterative times $I = 50$.

The number of topics in every temporal window is shown in Table 2.

Table 2. Topics numbers

Temporal series	September 1-15	September 16-30	October 1-15	October 16-30
Numbers	30	45	40	35

Then the K-means clustering algorithm is used to determine hot topics. Among the clustering parameters, K is 150, equal to the number of topics produced by LDA model on four temporal windows. There are two apparent topics among the clustering result in Figure 3. After referencing the document data, the two public opinion themes are respectively Topic 1 "Child Trafficking" and topic 2 "Temple One".

Since the number of hot topics is determined as 2, the temporal heat value model is applied to calculating topic heat value in every temporal window, of course $K = 2$. In the same way, the Gibbs sampling parameters are set as $\alpha^t = 50/T$ and $\beta^t = 0.01$, iterative times $I = 50$. According to the correlation information between one topic and its characteristic words, the top 10 characteristic words are relatively close to the topic, but in order to distinguish the difference of the topics more accurately, the top 15 characteristic words are extracted to represent each topic.

The heat values of each topic in every temporal windows are computed on the basis of Formula (5), and Figure 4 shows the result.

In Figure 4, the heat change of topic 1 on the 4 temporal windows is in accordance with the law of development of people's livelihood event – which consists of occurrence,

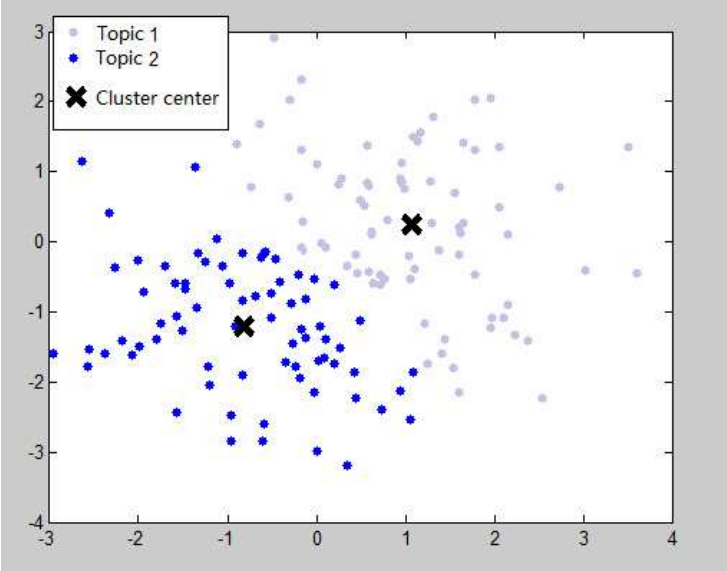

FIGURE 3. K-means clustering result

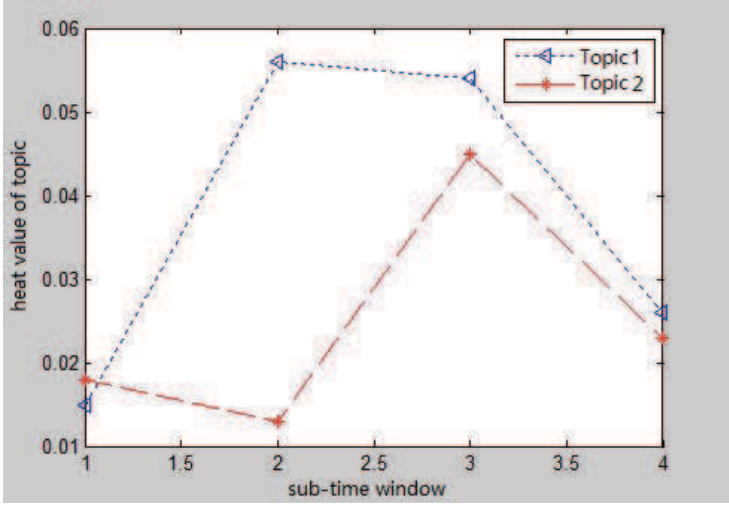

FIGURE 4. Changes of heat values

climax, continuous and fade. After the emergence of child trafficking, large number of users rapidly concern and forward microblogs about this topic. During a period of time after being peaked in popularity, the heat value of topic 1 is slightly down but still high, which confirms the high influence of the "Child Trafficking" event. Until some cases solved, the heat of topic 1 fades but is still higher than it appears, since there are many subsequent topics focusing on such as "punishment given". However, the science public events like "Temple One" will not receive much attention, even the heat will drop to the lowest point before new milestones occur. The hot topic popularity trend diagram scientifically shows the popularity of the two events in the period of emergence, development and recession, which is for public opinion monitoring departments to analyze various types of public events.

4.2. **Experiment evaluation.** The proposed model is compared with two existing methods: one is LDA model without label t, and the other is the R model in which the characteristic words are extracted from word set randomly. First, missing rate and false alarm rate are taken as indexes to evaluate the performance of the t-LDA model in TMT model on topic detection. For LDA model, initial K is 150, $\alpha^t = 50/T$ and $\beta^t = 0.01$, and the

number of iterations is 50. The experimental results are shown in Figure 5. Obviously, as the first step of TMT model, the t-LDA model generates a lower missing rate and a lower false alarm rate on topic detection, which means a more perfect detection result.

Secondly, F-measure is used to evaluate the clustering performance of TMT model. Formula (6) defines the F-measure.

$$F = \sum_i \frac{n_i}{n} \max\{f(i,r)\} \tag{6}$$

Among them, n represents the number of all test documents, and $f(i,r)$ means the F-measure of the cluster r and the predefined category i, which is computed as Formula (7).

$$f(i,r) = \frac{2 \times recall(i,r) \times precision(i,r)}{recall(i,r) + precision(i,r)} \tag{7}$$

$recall(i,r)$ and $precision(i,r)$ represent recall and precision respectively.

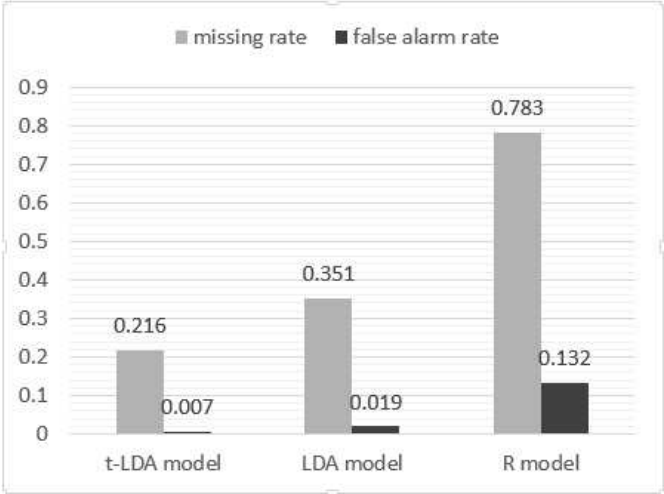

FIGURE 5. Comparison on topic detection

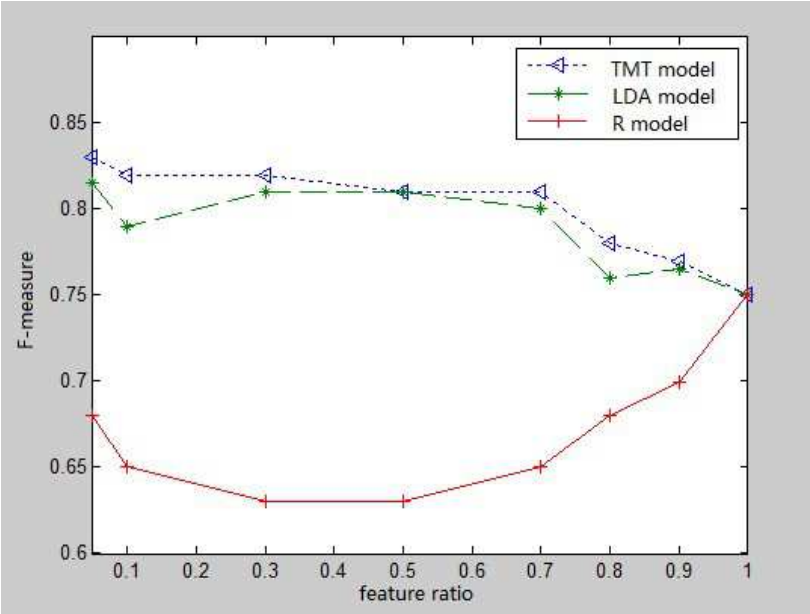

FIGURE 6. Comparison on F-measures

The K-means clustering algorithm is executed respectively based on the topic word sets got by the three methods mentioned above. Calculate recall and precision, and then get F-measure according to Formula (6). The result is shown as Figure 6. The comparison results show that TMT model gets a higher F-measure at any feature ratio, which means a better performance on clustering results.

5. **Conclusions.** In this paper, the TMT model is proposed for hot topic detection and analysis in microblog. The model is based on the fact that all topics have temporal feature. To make the process of topic detection more robust, the whole temporal period is discretized to several temporal windows, and the LDA model is combined with temporal label t so as to detect topics more precisely. To find hot topics, K-means clustering is applied on the detected topics. The proposed hot topic detection model is compared with two typical non-temporal topic detection methods. Temporal feature improves the accuracy of latent topic detection, and overcomes the problem arising from the sparseness of clustering data.

For future work, the identity of the topics could be considered in clustering process, since the evolution relationships between topics would be more complete with both identity and relevance. Also, it is worth unearthing other factors to enhance the performance of the TMT model, such as the sentiment of microblogs, the user of microblogs and post location of microblogs.

Acknowledgment. The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have greatly improved the presentation.

REFERENCES

[1] R. Li, K. H. Lei, R. Khadiwala et al., TEDAS: A twitter-based event detection and analysis system, *IEEE the 28th International Conference on Data Engineering*, vol.41, pp.1273-1276, 2012.

[2] D. Maynard, K. Bontcheva and D. Rout, Challenges in developing opinion mining tools for social media, *Proc. of the @NLP can u tag #user_generated_content?! workshop at LREC'12*, pp.15-22, 2012.

[3] M. Dredze, How social media will change public health, *IEEE Intelligent Systems*, vol.27, no.4, pp.81-84, 2012.

[4] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol.3, pp.993-1022, 2003.

[5] X. Yan, J. Guo, Y. Lan et al., A biterm topic model for short texts, *Proc. of the 22nd International Conference on World Wide Web*, pp.1445-1456, 2013.

[6] X. Wang, F. Zhu, J. Jiang et al., Real time event detection in twitter, *International Conference on Web-Age Information Management*, pp.502-513, 2013.

[7] B. Huang, Y. Yang, A. Mahmood et al., Microblog topic detection based on LDA model and single-pass clustering, *Rough Sets and Current Trends in Computing*, Springer Berlin Heidelberg, pp.166-171, 2012.

[8] G. Ifrim, B. Shi and I. Brigadir, Event detection in twitter using aggressive filtering and hierarchical tweet clustering, *Proc. of Snow*, 2014.

[9] Y. Chen, H. Amiri, Z. Li et al., Emerging topic detection for organizations from microblogs, *International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp.43-52, 2013.

[10] G. Stilo and P. Velardi, Efficient temporal mining of micro-blog texts and its application to event discovery, *Data Mining & Knowledge Discovery*, pp.1-31, 2015.

[11] D. Blei, L. Carin and D. Dunson, Probabilistic topic models, *IEEE Signal Processing Magazine*, vol.55, no.4, pp.55-65, 2010.